entropy

MDPI

Article

Entropies of Weighted Sums in Cyclic Groups and an Application to Polar Codes

Emmanuel Abbe [1,2], Jiange Li [3,4] and Mokshay Madiman [3,*]

[1] Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544-1000, USA; eabbe@princeton.edu

[2] Department of Electrical Engineering, Princeton University, Princeton, NJ 08544-1000, USA

[3] Department of Mathematical Sciences, University of Delaware, Newark, DE 19716, USA; lijiange@udel.edu

[4] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

* Correspondence: madiman@udel.edu; Tel.: +1-302-831-1865

Received: 14 February 2017; Accepted: 5 April 2017; Published: 7 September 2017

Abstract: In this note, the following basic question is explored: in a cyclic group, how are the Shannon entropies of the sum and difference of i.i.d. random variables related to each other? For the integer group, we show that they can differ by any real number additively, but not too much multiplicatively; on the other hand, for $\mathbb{Z}/3\mathbb{Z}$, the entropy of the difference is always at least as large as that of the sum. These results are closely related to the study of more-sums-than-differences (i.e., MSTD) sets in additive combinatorics. We also investigate polar codes for q-ary input channels using non-canonical kernels to construct the generator matrix and present applications of our results to constructing polar codes with significantly improved error probability compared to the canonical construction.

Keywords: entropy; more-sums-than-differences set; polar code

1. Introduction

For a discrete random variable X supported on a countable set A, its Shannon entropy $H(X)$ is defined to be:

$$H(X) = -\sum_{x\in A} \mathbb{P}(X = x) \log \mathbb{P}(X = x). \tag{1}$$

The Shannon entropy can be thought of as the logarithm of the effective cardinality of the support of X; the justification for this interpretation comes from the fact that when the alphabet A is finite, $H(X) \leq \log |A|$, with equality if and only if X is uniformly distributed on A. This suggests an informal parallelism between entropy inequalities and set cardinality inequalities that has been extensively explored for projections of subsets of Cartesian product sets (see, e.g., [1] for a review of these and their applications to combinatorics) and, more recently, for sums of subsets of a group that are of great interest in the area of additive combinatorics [2]. For two finite subsets A, B of an abelian group, the sumset $A + B$ and difference set $A - B$ are defined by:

$$A + B := \{a + b : a \in A, b \in B\},$$

and:

$$A - B := \{a - b : a \in A, b \in B\}.$$

In the trivial bound $\max\{|A|, |B|\} \le |A \pm B| \le |A||B|$, replacing the sets A, B by independent discrete random variables X, Y and replacing the log-cardinality of each set by the Shannon entropy, one obtains the entropy analogue:

$$\max\{H(X), H(Y)\} \le H(X \pm Y) \le H(X) + H(Y). \tag{2}$$

This is, of course, an analogy, but not a proof; however, the inequality (2) can be seen to be true from the elementary properties of entropy.

First identified by Ruzsa [3], this connection between entropy inequalities and cardinality inequalities in additive combinatorics has been studied extensively in the last few years. Useful tools in additive combinatorics have been developed in the entropy setting, such as Plünnecke–Ruzsa inequalities by Madiman, Marcus and Tetali [4,5] and Freiman–Ruzsa and Balog–Szemerédi–Gowers theorems by Tao [6]. Much more work has also recently emerged on related topics, such as efforts towards an entropy version of the Cauchy–Davenport inequality [7–10], an entropy analogue of the doubling-difference inequality [11] and applications of additive combinatorics in information theory [12–17]. Some results have also been extended from discrete abelian groups to locally compact abelian groups [18,19], with entropy being defined as an integral with respect to the Haar measure. In the particular case of the additive group $\mathbb{R}^d$, there continues to be a connection (see, e.g., [20–23]) between entropy inequalities for random variables and "size" inequalities for sumsets, except that size is taken to be the volume (Lebesgue measure) of a set rather than its cardinality. Entropy inequalities for sums of discrete random variables also have implications for probabilistic limit theorems (see, e.g., [24–27]), although this is not a direction we explore in this paper.

In an abelian group, since addition is commutative while subtraction is not, two generic elements generate one sum, but two differences. Likely motivated by this observation, the following conjecture (attributed to Conway) is contained in [28] (Section VI, Problem 7):

> "Let $A = \{a_1, a_2, \dots, a_N\}$ be a finite set of integers, and define $A + A = \{a_i + a_j : 1 \le i, j \le N\}$ and $A - A = \{a_i - a_j : 1 \le i, j \le N\}$. Prove that $A - A$ always has more members than $A + A$, unless A is symmetric about 0."

According to [29], Conway denied having conjectured the patently false statement about equality; apparently his original conjecture was that $A - A$ always has at least as many elements as $A + A$. However, that is not always the case. In 1969, Marica [30] showed that the conjecture is false by exhibiting the set $A = \{1, 2, 3, 5, 8, 9, 13, 15, 16\}$, for which $A + A$ has 30 elements and $A - A$ has 29 elements. Such a set is called an MSTD (more-sums-than-differences) set. According to Nathanson [31], Conway himself had already found the MSTD set $\{0, 2, 3, 4, 7, 11, 12, 14\}$ in the late 1960s, thus disproving his own conjecture. Subsequently, Stein [32] showed that one can construct sets A for which the ratio $|A - A|/|A + A|$ is as close to zero or as large as we please; apart from his own proof, he observed that such constructions also follow by adapting arguments in an earlier work of Piccard [33] that focused on the Lebesgue measure of $A + A$ and $A - A$ for subsets A of $\mathbb{R}$. A stream of recent papers aims to quantify how rare or frequent MSTD sets are (see, e.g., [34,35] for work on the integers and [36] for finite abelian groups more generally) or try to provide denser constructions of infinite families of MSTD sets (see, e.g., [37,38]); however, these are not directions we will explore in this note.

Since convolutions of uniforms are always distributed on the sumset of the supports, but are typically not uniform distributions, it is not immediately obvious from the Conway and Marica constructions whether there exist i.i.d. random variables X and Y such that $H(X + Y) > H(X - Y)$. The purpose of this note is to explore this and related questions. For example, one natural related question to ask is for some description of the coefficient $\lambda \in \{1, \dots, |G|\}$ that maximizes $H(X + \lambda Y)$ for X, Y drawn i.i.d. from some distribution in G; restricting the choice of coefficients to $\{+1, -1\}$ would correspond to the sum-difference question. This question is motivated by applications to the class of polar codes, which is a very promising class of codes that has attracted much recent attention in information and coding theory. Specifically, we show that over $\mathbb{F}_q$, the "spread" of the polar martingale can be significantly enlarged by using optimized kernels rather than the original kernel $\left[\begin{smallmatrix}1 & 0\\ 1 & 1\end{smallmatrix}\right]$. In some cases, this leads to significant improvements on the error probability of polar codes, even at low block lengths like 1024. We also consider additive noise channels and show that the improvement is particularly significant when the noise distribution is concentrated on a "small" support.

This note is organized as follows. In Section 2.1, we show that entropies of sums (of i.i.d. random variables) are never greater than entropies of differences for random variables taking values in the cyclic group $\mathbb{Z}/3\mathbb{Z}$; however, this fails for larger groups, and in particular, we show that there always exist distributions on finite cyclic groups of order at least 21 such that $H(X+Y) > H(X-Y)$. In Sections 2.2 and 2.3, we explore more quantitative questions; that is, we ask not only what the ordering of $H(X+Y)$ and $H(X-Y)$ may be, but how different these can be in either direction; the finding here is that on $\mathbb{Z}$, these can differ by arbitrarily large amounts additively, but not too much multiplicatively. Finally, in Section 3, we explore the question about entropies of weighted sums mentioned at the end of the previous paragraph and describe the applications to polar codes, as well.

2. Comparing Entropies of Sums and Differences

2.1. Basic Examples

We start by considering the smallest group in which the sum and difference are distinct, namely $\mathbb{Z}/3\mathbb{Z}$. Let $p = (p_0, p_1, p_2)$ be a probability distribution on $\mathbb{Z}/3\mathbb{Z}$, and let $H(p)$ be its Shannon entropy. We denote by $\|p - U\|_2$ the Euclidean distance between p and the uniform distribution $U = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. For any fixed $0 \le t \le \log 3$, the following lemma verifies the "triangular" shape of the entropy circle $H(p) = t$.

Lemma 1. *Let p be a probability distribution on the entropy circle $H(p) = t$ such that $p_0 \ge p_1 \ge p_2$. Then, the distance $\|p - U\|_2$ is an increasing function of p_0.*

Proof. If $t = 0$, then p has to be the deterministic distribution $(1, 0, 0)$. In this case, we have $\|p - U\|_2 = \sqrt{2/3}$. If $t = \log 3$, we have $p = U$ and $\|p - U\|_2 = 0$. In the following, we may assume that $0 < t < \log 3$. The condition $p_0 + p_1 + p_2 = 1$ yields:

$$1 + \frac{dp_1}{dp_0} + \frac{dp_2}{dp_0} = 0. \tag{3}$$

The entropy identity $H(p) = t$ implies:

$$(\log p_0 + 1) + (\log p_1 + 1)\frac{dp_1}{dp_0} + (\log p_2 + 1)\frac{dp_2}{dp_0} = 0. \tag{4}$$

The above two identities give us that:

$$\frac{dp_1}{dp_0} = \frac{\log p_0 - \log p_2}{\log p_2 - \log p_1} \tag{5}$$

and:

$$\frac{dp_2}{dp_0} = \frac{\log p_0 - \log p_1}{\log p_1 - \log p_2}. \tag{6}$$

Using Identities (3), (5) and (6), we have:

$$\begin{aligned}
\frac{1}{2}\cdot\frac{d}{dp_0}\|p-U\|_2 &= \sum_{i=0}^{2}\left(p_i-\frac{1}{3}\right)\frac{dp_i}{dp_0}\\
&= p_0+p_1\frac{\log p_0-\log p_2}{\log p_2-\log p_1}+p_2\frac{\log p_0-\log p_1}{\log p_1-\log p_2}\\
&= (p_0-p_1)\frac{\log p_0-\log p_2}{\log p_1-\log p_2}-(p_0-p_2)\frac{\log p_0-\log p_1}{\log p_1-\log p_2}\\
&= \frac{(p_0-p_1)(p_0-p_2)}{\log p_1-\log p_2}\left(\frac{\log p_0-\log p_2}{p_0-p_2}-\frac{\log p_0-\log p_1}{p_0-p_1}\right)\\
&\geq 0.
\end{aligned}$$

The last inequality follows from the assumption that $p_0 \geq p_1 \geq p_2$ and the concavity of the logarithmic function. □

Now, we can show that the entropy of the sum of two i.i.d. random variables taking values in $\mathbb{Z}/3\mathbb{Z}$ can never exceed the entropy of their difference. We use basic facts about the Fourier transform on finite groups, which can be found, e.g., in [39].

Theorem 1. *Let X, Y be i.i.d. random variables taking values in $\mathbb{Z}/3\mathbb{Z}$, then we have:*

$$H(X+Y) \leq H(X-Y). \tag{7}$$

Proof. Let $p = (p_0, p_1, p_2)$ be the distribution of X. Since Y is an independent copy of X, we can see that $-Y$ has distribution $q = (p_0, p_2, p_1)$. Then, the distributions of $X+Y$ and $X-Y$ can be written as $p \star p$ and $p \star q$, respectively, where "$\star$" is the convolution operation. Let $\widehat{p} = (\widehat{p}_0, \widehat{p}_1, \widehat{p}_2)$ be the Fourier transform of p with Fourier coefficients defined by:

$$\widehat{p}_j = \sum_{k=0}^{2} p_k e^{-i2\pi jk/3}, \;\; j = 0, 1, 2.$$

One basic property of the Fourier transform asserts that:

$$\widehat{q}_j = \overline{\widehat{p}_j}, \tag{8}$$

where $\overline{\widehat{p}_j}$ is is the conjugate of $\widehat{p}_j$. We also have:

$$(\widehat{p \star q})_j = \widehat{p}_j \cdot \widehat{q}_j, \tag{9}$$

which holds for general distributions q. The Parseval–Plancherel identity says:

$$\|\widehat{p}\|_2^2 = 3\|p\|_2^2. \tag{10}$$

Using the identities (8)–(10), we have:

$$\|p \star p\|_2 = \|p \star q\|_2,$$

which implies:

$$\|p \star p - U\|_2 = \|p \star q - U\|_2.$$

It is not hard to see that $X-Y$ is symmetric with $(p \star q)_0 \geq (p \star q)_1 = (p \star q)_2$. Using Lemma 1, we can see that the entropy circle passing through $p \star q$ lies inside the Euclidean circle centered at U

with radius $\|p \star q - U\|_2$. Thus, the distribution $p \star p$ is on an entropy circle with entropy not greater than $H(p \star q)$. Then, we have the desired statement. □

The property in Theorem 1 fails to hold for larger cyclic groups; we demonstrate this by discussing three specific examples of i.i.d. random variables X, Y such that the entropy of their sum is larger than the entropy of their difference.

1. For Conway's MSTD set $A = \{0,2,3,4,7,11,12,14\}$, we have $|A+A| = 26$ and $|A-A| = 25$. Let X, Y be independent random variables uniformly distributed on A. Straightforward calculations show that:
$$H(X+Y) - H(X-Y) = \frac{1}{64}\log\frac{282429536481}{215886856192} > 0.$$
2. The second example is based on the set $A = \{0,1,3,4,5,6,7,10\}$ with $|A+A| = |A-A| = 19$. Let X, Y be independent random variables uniformly distributed on A. Then, we have:
$$H(X+Y) - H(X-Y) = \frac{1}{64}\log\frac{5^{10}\times 8^{10}}{3^6\times 7^7} > 0.$$
3. The group $\mathbb{Z}/12\mathbb{Z}$ is the smallest cyclic group that contains an MSTD set. Let $A = \{0,1,2,4,5,9\}$. It is easy to check that A is an MSTD set since $A + A = \mathbb{Z}/12\mathbb{Z}$ and $A - A = (\mathbb{Z}/12\mathbb{Z})\backslash\{6\}$. We let X, Y be independent random variables uniformly distributed on A. Then, we have:
$$H(X+Y) - H(X-Y) = \frac{1}{36}\log\frac{3^{34}}{20^{10}} > 0.$$

Remark 1. *Applying linear transformations, we can get infinitely many MSTD sets of $\mathbb{Z}$ from Conway's MSTD set. Correspondingly, one can get as many "MSTD" random variables as one pleases. Thus, MSTD sets are useful in the construction of "MSTD" random variables; however, we can also construct "MSTD" random variables supported on non-MSTD sets as shown by the second example.*

Remark 2. *Hegarty [40] proved that there is no MSTD set in $\mathbb{Z}$ of size seven, and up to linear transformations, Conway's set is the unique MSTD set in $\mathbb{Z}$ of size eight. We do not know the smallest support of "MSTD" random variables taking values in $\mathbb{Z}$, although eight is clearly an upper bound.*

Remark 3. *We also do not know the smallest m such that there exist "MSTD" random variables taking values in $\mathbb{Z}/m\mathbb{Z}$; however, the third example shows that this m cannot be greater than* 12.

2.2. Achievable Differences

We first briefly introduce the construction of Stein [32] of finite subsets $A_k \subset \mathbb{Z}$ such that the ratio $|A_k - A_k|/|A_k + A_k|$ can be arbitrarily large or small when k is large. Using this construction, we will give an alternate proof of the result of Lapidoth and Pete [12], which asserts that $H(X-Y)$ can exceed $H(X+Y)$ by an arbitrarily large amount.

Let $A, B \subset \mathbb{Z}$ be two finite subsets. Suppose that the gap between any two consecutive elements of B is sufficiently large. For any $b \in B$, the set $b + A$ represents a relatively small fluctuation around b. Large gaps between elements of B will imply that $(b + A) \cap (b' + A) = \emptyset$ for distinct $b, b' \in B$. Then, we will have $|A + B| = |A||B|$. For $m \in \mathbb{Z}$ large, this argument implies that $|A + m \cdot A| = |A|^2$, where $m \cdot A := \{ma : a \in A\}$. Therefore, the following equations hold simultaneously for sufficiently large $m_0 \in \mathbb{Z}$ (which depends on A, $A - A$ and $A + A$):

$$|A + m_0 \cdot A| = |A|^2,$$

$$|(A + m_0 \cdot A) - (A + m_0 \cdot A)| = |(A - A) + m_0 \cdot (A - A)| = |A - A|^2,$$

and:

$$|(A+m_0\cdot A)+(A+m_0\cdot A)|=|A+A|^2.$$

Repeating this argument, we can get a sequence of sets A_k, defined by:

$$A_k=A_{k-1}+m_{k-1}A_{k-1}, \tag{11}$$

where $A_0=A$, $m_{k-1}\in\mathbb{Z}$ sufficiently large, with the following properties:

$$|A_k|=|A|^{2k},\ \ |A_k\pm A_k|=|A\pm A|^{2k}. \tag{12}$$

Now, we are ready to reprove the result of Lapidoth and Pete [12].

Theorem 2 ([12])**.** *For any $M>0$, there exist i.i.d. $\mathbb{Z}$-valued random variables X,Y with finite entropy such that:*

$$H(X-Y)-H(X+Y)>M.$$

Proof. Recall the following basic property of Shannon entropy:

$$0\le H(X)\le\log|\text{range of }X|. \tag{13}$$

We let X_k,Y_k be independent random variables uniformly distributed on the set A_k obtained by the iteration Equation (11). Using the right-hand side of (13) and the properties given by (12), we have:

$$H(X_k+Y_k)\le\log|A_k+A_k|=2k\log|A+A|. \tag{14}$$

Since X_k,Y_k are independent and uniform on A_k, for all $x\in A_k-A_k$, we have:

$$\mathbb{P}(X_k-Y_k=x)\ge|A_k|^{-2}.$$

Notice the fact that $-t\log t$ is increasing over $(0,1/e)$. When k is large enough, we have:

$$\begin{aligned}H(X_k-Y_k) &\ge \frac{|A_k-A_k|}{|A_k|^2}\log|A_k|^2\\ &= 4k\log|A|\left(\frac{|A-A|}{|A|^2}\right)^{2k}.\end{aligned} \tag{15}$$

For any $k\in\mathbb{Z}^+$, we can always find a set $A\subset\mathbb{Z}$ with k^2 elements such that the set $A-A$ achieves the possible maximal cardinality,

$$|A|=k^2,\ |A-A|=|A|^2-|A|+1. \tag{16}$$

Combining (14), (16) and the trivial bound:

$$|A+A|\le\frac{|A|(|A|+1)}{2},$$

we have that for k large:

$$\begin{aligned}H(X_k+Y_k) &\le 2k\log\frac{|A|(|A|+1)}{2}\\ &= 8k\log k-2k\log 2+2k\log(1+k^{-2})\\ &= 8k\log k-2k\log 2+o(1).\end{aligned}$$

Combining (15) and (16), we have:

$$\begin{aligned} H(X_k - Y_k) &\geq 8k\log k\left(1 - k^{-2} + k^{-4}\right)^{2k} \\ &= 8k\log k\exp(2k(-k^{-2} + O(k^{-4}))) \\ &= 8k\log k(1 - 2k^{-1} + O(k^{-2})) \\ &= 8k\log k - 16\log k + o(1). \end{aligned}$$

Therefore, we have:

$$H(X_k - Y_k) - H(X_k + Y_k) = 2k\log 2 - 16\log k + o(1).$$

Then, the statement follows from that k can be arbitrarily large. □

We observe that the following complementary result is also true.

Theorem 3. *For any $M > 0$, there exist i.i.d. $\mathbb{Z}$-valued random variables X, Y with finite entropy such that:*

$$H(X + Y) - H(X - Y) > M.$$

Remark 4. *The previous argument cannot be used to prove this result. If we proceed with the same argument, we will see that the lower bound of $H(X_k + Y_k)$ similar to (15) will be really bad. The reason is that:*

$$\left(\frac{|A + A|}{|A|^2}\right)^{2k} \to 0$$

exponentially fast. Both Theorems 2 and 3 can be proven using a probabilistic construction of Ruzsa [41] on the existence of large additive sets A with $|A - A|$ very close to the maximal value $|A|^2$, but $|A + A| \leq n^{2-c}$ for some explicit absolute constant $c > 0$; and similarly, with the roles of $A - A$ and $A + A$ reversed.

In fact, we have the following stronger result.

Theorem 4. *For any $M \in \mathbb{R}$, there exist i.i.d. $\mathbb{Z}$-valued random variables X, Y with finite entropy such that:*

$$H(X + Y) - H(X - Y) = M.$$

Proof. Let X be a random variable taking values in $\{0, 1, \cdots, n-1\} \subset \mathbb{Z}$. Then, $H(X + Y) - H(X - Y)$ is a continuous function of the probability mass function of X, which consists of n real variables. We can assume that n is large enough if necessary. From the discussion in Section 2.1, we know that this function can take both positive and negative values (for instance, Theorem 1 implies that a binary distribution can give us negative difference, and the uniform distribution on Conway's MSTD set will yield a positive difference). Since the function is continuous, the intermediate value theorem implies that its range must contain an open interval (a, b) with $a < 0 < b$. Let $X_1, \cdots, X_k$ be k independent copies of X, and we define $X' = (X_1, \cdots, X_k)$. Let Y' be an independent copy of X'. Then, we have:

$$H(X' + Y') - H(X' - Y') = k[H(X + Y) - H(X - Y)].$$

The range of $H(X' + Y') - H(X' - Y')$ will contain (ka, kb). This difference can take any real number since k can be arbitrarily large. The random variables X', Y' take finite values of $\mathbb{Z}^k$. Using the linear transformation $(x_1, \cdots, x_k) \to x_1 + dx_2 + \cdots + d^{k-1}x_k$, we can map X, Y to $\mathbb{Z}$-valued random variables. This map preserves entropy as d is large enough. Therefore, these $\mathbb{Z}$-valued random variables will have the desired property. □

Recall that, for a real-valued random variable X with the density function $f(x)$, the differential entropy $h(X)$ is defined by:

$$h(X) = -\mathbb{E}\log f(X) = -\int_{\mathbb{R}} f(x)\log f(x)dx. \tag{17}$$

Theorem 5. *For any $M \in \mathbb{R}$, there exist i.i.d. real-valued random variables X, Y with finite differential entropy, such that:*

$$h(X+Y) - h(X-Y) = M. \tag{18}$$

Proof. From Theorem 4, we know that there exist $\mathbb{Z}$-valued random variables X', Y' with the desired property. Let U, V be independent random variables uniformly distributed on $(-1/4, 1/4)$, which are also independent of (X', Y'). Then, we define $X = X' + U$ and $Y = Y' + V$. Elementary calculations will show that:

$$h(X+Y) = H(X'+Y') + h(U+V),$$

and:

$$h(X-Y) = H(X'-Y') + h(U-V).$$

Since U, V are symmetric, $U+V$ and $U-V$ have the same distribution. Therefore, we have:

$$h(X+Y) - h(X-Y) = H(X'+Y') - H(X'-Y').$$

Then, the theorem follows. □

Remark 5. *In the set cardinality setting, Nathanson [42] raised the question: what are the possible values of $|A+A| - |A-A|$ for finite subsets $A \subset \mathbb{Z}$? Martin and O'Bryant [34] proved that for any $k \in \mathbb{Z}$, there exists A such that $|A+A| - |A-A| = k$; this was also independently obtained by Hegarty [40].*

Remark 6. *It is interesting to contrast Theorem 5 with the observation in Remark 7.1 of [43] that for i.i.d. real-valued random variables X, Y, $h_\infty(X-Y) - h_\infty(X+Y) \le \log 2$, where h_∞ denotes the Rényi differential entropy of order infinity (i.e., if X has density f, $h_\infty(X) = -\log \operatorname{ess\,sup}_x f(x)$).*

2.3. Entropy Analogue of the Freiman–Pigarev Inequality

We proved that the entropies of the sum and difference of two i.i.d. random variables can differ by an arbitrary amount additively. However, we will show that they do not differ too much multiplicatively.

In additive combinatorics, for a finite additive set A, the doubling constant $\sigma[A]$ is defined as:

$$\sigma[A] = \frac{|A+A|}{|A|}. \tag{19}$$

Similarly, the difference constant $\delta[A]$ is defined by:

$$\delta[A] = \frac{|A-A|}{|A|}. \tag{20}$$

It was first observed by Ruzsa [44] that:

$$\delta[A]^{1/2} \le \sigma[A] \le \delta[A]^3. \tag{21}$$

The upper bound can be improved down to $\delta[A]^2$ using Plünnecke inequalities. Thus, a finite additive set has a small doubling constant if and only if its difference constant is also small. In the entropy setting, we have:

$$\frac{1}{2} \leq \frac{H(X+Y)-H(X)}{H(X-Y)-H(X)} \leq 2 \tag{22}$$

for i.i.d. random variables X, Y. The upper bound was proven by Madiman [13], and the lower bound was proven independently by Ruzsa [3] and Tao [6]. The inequalities also hold for differential entropy [11,18] and in fact for entropy with respect to the Haar measure on any locally compact abelian group [19]. In other words, after subtraction of $H(X)$, the entropies of the sum and the difference of two i.i.d. random variables are not too different. We observe that the entropy version (22) of the doubling-difference inequality implies the entropy analogue of the following result proven by Freiman and Pigarev [45]:

$$|A-A|^{3/4} \leq |A+A| \leq |A-A|^{4/3}. \tag{23}$$

Theorem 6. *Let X, Y be i.i.d. discrete random variables with finite entropy, then we have:*

$$\frac{3}{4} < \frac{H(X+Y)}{H(X-Y)} < \frac{4}{3}. \tag{24}$$

Proof. The basic fact of Shannon entropy (2) implies that $H(X+Y) = 0$ if and only if $H(X-Y) = 0$. In this case, the above theorem is true if we define $0/0 = 1$. Therefore, we assume that neither $H(X+Y)$ nor $H(X-Y)$ is zero. For the upper bound, we have:

$$\begin{aligned} \frac{H(X+Y)}{H(X-Y)} &= \frac{H(X+Y)}{H(X-Y)-H(X)+H(X)} \\ &\leq \frac{H(X+Y)}{(H(X+Y)-H(X))/2+H(X)} \\ &= \frac{2H(X+Y)}{H(X+Y)+H(X)} \\ &< \frac{4}{3} \end{aligned}$$

The second step follows from the upper bound in (22) and the fact that Shannon entropy is non-negative. The last step uses the right-hand side of (2) and the fact that, in the i.i.d. case, "=" of the upper bound happens only when X takes on a single value, i.e., $H(X) = 0$. The lower bound can be proven in a similar way. □

Remark 7. *It is unknown if the inequality (22) is the best possible. Suppose that, for some $\alpha \in (1, 2)$, we have:*

$$\alpha^{-1} \leq \frac{H(X+Y)-H(X)}{H(X-Y)-H(X)} \leq \alpha.$$

Using the same argument, the above theorem can be improved to:

$$\frac{\alpha+1}{2\alpha} < \frac{H(X+Y)}{H(X-Y)} < \frac{2\alpha}{\alpha+1}.$$

Remark 8. *The above theorem does not hold for continuous random variables. For example, let X be an exponential random variable with parameter λ and Y be an independent copy of X. Then, $X+Y$ satisfies the Gamma distribution $\Gamma(2, \lambda^{-1})$ with the differential entropy:*

$$h(X+Y) = 1 + \gamma - \log\lambda \approx 1.577 - \log\lambda,$$

where γ is Euler's constant. On the other hand, $X - Y$ has the Laplace distribution Laplace$(0, \lambda^{-1})$ with the differential entropy:

$$h(X-Y) = 1 + \log 2 - \log\lambda \approx 1.693 - \log\lambda.$$

We can see that:

$$\lim_{\lambda\to(2e)^+} \frac{h(X+Y)}{h(X-Y)} = \infty,$$

and:

$$\lim_{\lambda\to(2e)^-} \frac{h(X+Y)}{h(X-Y)} = -\infty.$$

3. Weighted Sums and Polar Codes

3.1. Polar Codes: Introduction

Polar codes, invented by Arıkan [46] in 2009, achieve the capacity of arbitrary binary-input symmetric discrete memoryless channels. Moreover, they have low encoding and decoding complexity and an explicit construction. Consequently, they have attracted a great deal of attention in recent years. In order to discuss polar codes more precisely, we now recall some standard terminology from information and coding theory.

As is standard practice in information theory, we use U^k to denote $(U_1, \dots, U_k)$ and $I(X;Y|Z)$ to denote the conditional mutual information between X and Y given Z, which is defined by:

$$I(X;Y|Z) = H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z).$$

It is well known and also trivial to see that the conditional entropy $H(X|Y)$, defined as the mean using the distribution of Y of $H(X|Y=y)$, satisfies the "chain rule" $H(Y) + H(X|Y) = H(X,Y)$, so that $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$. The mutual information between X and Y, namely $I(X;Y) = H(X) - H(X|Y)$, emerges in the case where there is no conditioning. In particular, $I(X;Y|Z) = 0$ if and only if X and Y are conditionally independent given Z. Furthermore, one also has the chain rule for mutual information, which states that $I(X;Y,Z) = I(X;Z) + I(X;Y|Z)$.

A major goal in coding theory is to obtain efficient codes that achieve the Shannon capacity on a discrete memoryless channel. A memoryless channel is defined first by a "one-shot" channel W, which is a stochastic kernel from an input alphabet $\mathcal{X}$ to an output alphabet $\mathcal{Y}$ (i.e., for each $x \in \mathcal{X}$, $W(\cdot|x)$ is a probability distribution on $\mathcal{Y}$), and the memoryless extension of W for length n vectors is defined by:

$$W^{(n)}(y^n|x^n) = \prod_{i=1}^{n} W(y_i|x_i), \quad x^n \in \mathcal{X}^n, y^n \in \mathcal{Y}^n. \tag{25}$$

To simplify the notation, one often makes a slight abuse of notation, writing $W^{(n)}$ as W.

A linear code of block length n on an alphabet $\mathcal{X} = \mathbb{F}$ (which must be a field) is a subspace of $\mathbb{F}^n$. The vectors in the subspace are often called the codewords. A linear code is equivalently defined by a generator matrix, i.e., a matrix with entries in the field whose rows form a basis for the code. If the dimension of the code is k and if G is a $k \times n$ generator matrix for the linear code, the codewords are given by the span of the rows of G, i.e., all multiplications uG where u is a $1 \times k$ vector over the field. We refer to [47,48] for a more detailed introduction to information and coding theory.

In polar codes, the generator matrix of block length n is obtained by deleting some rows of the matrix $G_n = \left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]^{\otimes \log_2 n}$. (If the channel is symmetric, the generator matrix is indeed obtained by deleting rows; otherwise, in addition to deleting rows, one may also have to translate the codewords, i.e., use an affine code.) Which rows to delete depends on the channel and the targeted error probability (or rate). For a symmetric discrete memoryless channel W, the rows to be deleted are indexed by:

$$\mathcal{B}_{\epsilon,n} := \{i \in [n] : I(U_i; Y^n U^{i-1}) \leq 1 - \epsilon\}, \tag{26}$$

where ϵ is a parameter governing the error probability, the vector U^n has i.i.d. components, which are uniform on the input alphabet, $X^n = U^n G_n$, and Y^n is the output of n independent uses of W when X^n is the input.

To see the purpose of the transform G_n, consider the case $n = 2$ first. Applying G_2 to the vector (U_1, U_2) yields:

$$\begin{aligned} X_1 &= U_1 + U_2, \\ X_2 &= U_2. \end{aligned}$$

Transmitting X_1 and X_2 on two independent uses of a binary input channel W leads to two output variables Y_1 and Y_2; recall that this means that Y_1 (or Y_2) is a random variable whose distribution is given by $W(\cdot|x)$ where x is the realization of X_1 (or X_2). If we look at the mutual information between the vectors $X^2 = (X_1, X_2)$ and $Y^2 = (Y_1, Y_2)$, since the pair of components (X_1, Y_1) and (X_2, Y_2) are mutually independent, the chain rule yields:

$$I(X^2; Y^2) = I(X_1; Y_1) + I(X_2; Y_2) = 2I(W), \tag{27}$$

where $I(W)$ is defined as the mutual information of the one-shot channel W with a uniformly-distributed input. Further, since the transformation G_2 is one-to-one and since the mutual information is clearly invariant under one-to-one transformations of its arguments (the mutual information depends only on the joint distribution of its arguments), we have that:

$$I(U^2; Y^2) = I(X^2; Y^2). \tag{28}$$

If we now apply the chain rule to the left-hand side of the previous equality, the dependencies in the components of U^2 obtained by mixing X^2 with G_2 lead this time to two different terms, namely,

$$I(U^2; Y^2) = I(U_1; Y^2) + I(U_2; Y^2, U_1). \tag{29}$$

Putting back (27)–(29) together, we have that:

$$I(W) = \frac{1}{2}\left(I(U_1; Y^2) + I(U_2; Y^2, U_1)\right). \tag{30}$$

Now, the above is interesting because the two terms in the right-hand side are precisely not equal. In fact, $I(U_2; Y^2, U_1)$ must be greater than its counter-part without the mixing of G_2, i.e., $I(U_2; Y^2, U_1) \geq I(X_2; Y_2) = I(W)$. To see this, note that:

$$\begin{aligned} I(U_2; Y^2, U_1) &= H(U_2) - H(U_2|Y^2, U_1) \\ &\geq H(U_2) - H(U_2|Y^2) \\ &= H(X_2) - H(X_2|Y_2) \\ &= I(X_2; Y_2) \end{aligned}$$

where the inequality above uses the fact that conditioning can only reduce entropy; hence, dropping the variable U_1 in $H(U_2|Y^2, U_1)$ can only increase the entropy. Further, one can check that besides for degenerated cases where W is deterministic or fully noisy (i.e., making input and output independent),

$I(U_2; Y^2, U_1)$ is strictly larger than $I(X_2; Y_2)$. Thus, the two terms in the right-hand side of (30) are respectively lesser and greater that $I(W)$, but they average out to the original amount $I(W)$.

In summary, out of two independent copies of the channel W, the transform G_2 allows us to create two new synthetic channels:

$$W^- : U_1 \to Y_1, Y_2$$
$$W^+ : U_2 \to Y_1, Y_2, U_1$$

that have respectively a worse and better mutual information:

$$I(W^-) \le I(W) \le I(W^+).$$

while overall preserving the total amount of mutual information:

$$I(W) = \frac{1}{2}(I(W^+) + I(W^-)).$$

The key use of the above phenomena is that if one wants to transmit only one bit (uniformly drawn), using W^+ rather than W leads to a lower error probability since the channel W^+ carries more information. One can then iterate this argument several times and hope to obtain a subset of channels of very high mutual information, on which bits can be reliably transmitted. After $\log_2 n$ iterations, one obtains the synthesized channels $U_i \mapsto (Y^n, U^{i-1})$. Thus, for a given number of information bits to be transmitted (i.e., for a given rate), one can select the channels with the largest mutual information to minimize the error probability. As explained in the next section, the phenomenon of polarization happens in the sense that as n tends to infinity, the synthesized channels have mutual information tending to either zero or one (besides for a vanishing fraction of exceptions). Hence, sending information bits through the high mutual information channels (equivalently, deleting rows of G_n corresponding to low mutual information channels) allows one to achieve communication rates as large as the mutual information of the original binary input channel. The construction extends to q-ary input alphabets when q is prime using the same matrix $G_n = \left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]^{\otimes \log_2 n}$, while carrying the operations over $\mathbb{F}_q$. (If q is a power of a prime and one uses modulo q operations, the polarization still occurs, but to multiple levels, as shown independently by [49,50].)

It is tempting to investigate what happens if one keeps the Kronecker structure of the generator matrix, but modifies the kernel $\left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]$. For binary input alphabets, there is no other interesting choice (up to equivalent permutations). In Mori and Tanaka [51], the error probability of non-binary polar codes constructed on the basis of Reed–Solomon matrices is calculated using numerical simulations on q-ary erasure channels. It is confirmed that 4-ary polar codes can have significantly better performance than binary polar codes. Our goal here is to investigate potential improvements at finite block length using modified kernels over $\mathbb{F}_q$. We propose to pick kernels not by optimizing the polar code exponent as in [51], but by maximizing the polar martingale spread. This connects to the object of study in this paper, as explained next. The resulting improvements are illustrated with numerical simulations.

3.2. Polar Martingale

In order to see that polarization happens, namely that:

$$\frac{1}{n}|\{i \in [n] : I(U_i; Y^n, U^{i-1}) \in (\epsilon, 1-\epsilon)\}| \to 0, \tag{31}$$

it is helpful to rely on a random process having a uniform measure on the possible realizations of $I(U_i; Y^n U^{i-1})$. Then, counting the number of such mutual information in $(\epsilon, 1-\epsilon)$ can be obtained by evaluating the probability that the process lies in this interval. The process is defined by taking

$\{B_n\}_{n\geq 1}$ to be i.i.d. random variables uniform on $\{-,+\}$, and the binary (or q-ary with q prime) random input channels $\{W_n,\ n\geq 0\}$ are defined by:

$$\begin{aligned} &W_0 := W, \\ &W_n := W_{n-1}^{B_n}, \quad \forall n\geq 1. \end{aligned} \tag{32}$$

Then, the polarization result can be expressed as:

$$\mathbb{P}\{I(W_n)\in(\epsilon,1-\epsilon)\}\to 0. \tag{33}$$

The process $I(W_n)$ is particularly handy as it is a bounded martingale with respect to the filtration B_n. This is a consequence of the balance equation derived in (30). Therefore, $I(W_n)$ converges almost surely, which means that almost surely, for any $\epsilon>0$ and n large enough, $|I(W_{n+1})-I(W_n)|=I(W_n^+)-I(W_n)<\epsilon$. Since for q-ary input channels (q prime), the only channels for which $I(W^+)-I(W)$ is arbitrarily small is when $I(W)$ is arbitrarily close to zero or one, the conclusion of polarization follows. The key point is that the martingale $I(W_n)$ is a random walk in $[0,1]$, and it is unstable at any point $I(W)\in(0,1)$ as it must move at least $I(W^+)-I(W)>0$ in this range. The plot of $I(W^+)-I(W)>0$ for different values of $I(W)$ is provided in Figure 1.

Thus, the larger the spread $I(W^+)-I(W)$, the more unstable the martingale is at non-extremal points and the faster it should converge to the extremes (i.e., polarized channels). To see why this is connected to the object of study of this paper, we need one more aspect about polar codes.

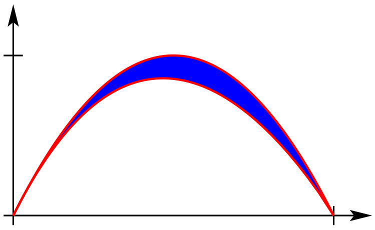

Figure 1. Plot of $I(W)$ (horizontal axis) vs. $I(W^+)-I(W)$ for all possible binary input channels (the tick on the horizontal axis is at one, and the tick on vertical axis is at $1/4$).

When considering channels that are "additive noise", polarization can be understood in terms of the noise process rather than the actual channels W_n. Consider for example the binary symmetric channel. When transmitting a codeword c^n on this channel, the output is $Y^n=c^n+Z^n$, where Z^n has i.i.d. Bernoulli components. The polar transform can then be carried over the noise Z^n. Since:

$$I(U_i;Y^nU^{i-1})=1-H((G_nZ^n)_i|(G_nZ^n)^{i-1}), \tag{34}$$

the mutual information of the polarized channels is directly obtained from the conditional entropies of the polarized noise vector G_nZ^n. The counterpart of this polarization phenomenon is called source polarization [52]. It is extended in [53] to multiple correlated sources. For $n=2$, the spread of the two conditional entropies is exactly given by $H(Z+Z')-H(Z)$, where Z,Z' are i.i.d. under the noise distribution. In Arıkan and Telatar [54], the rate of convergence of the polar martingale is studied as a function of the block length. Our goal here is to investigate the performance at finite block length, motivated by maximizing the spread at block length $n=1$. When considering non-binary polar codes, that spread is governed by the entropy of a linear combination of i.i.d. variables. Preliminary results on this approach were presented in [55].

One should also mention that several works have investigated the scaling law of polar codes. In particular, the scaling exponent is characterized in [56,57] and is shown to be between three and four (in contrast to an exponent of two for random codes), with further details available in [58]. Other works have also studied the effect of using kernels that have dimension greater than two, such as in [59–61]. Such approaches allow one to achieve a probability of error that decays faster than exponential in the square root of the block length, in fact almost exponential for arbitrary large kernels, but to the expense of a significant increase in complexity (leaving dimension two the most relevant dimension for practical applications).

3.3. Kernels with Maximal Spread

Being interested in the performance of polar codes at finite block length, we start with the optimization of the kernel matrix over $\mathbb{F}_q$ of block length $n = 2$. Namely, we investigate the following optimization problem:

$$K^*(W) = \arg \max_{K \in M_2(\mathbb{F}_q)} I(W^+(W, K)), \tag{35}$$

where $W^+(W, K)$ is the channel $u_2 \mapsto Y_1 Y_2 u_1$ and (Y_1, Y_2) are the output of two independent uses of W when $(x_1, x_2) = (u_1, u_2)K$ are the inputs. We call K^* the two-optimal kernel for W.

A general kernel is a 2×2 invertible matrix over $\mathbb{F}_q$. Let $K = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ be such a matrix, and let (U_1, U_2) be i.i.d. under μ over $\mathbb{F}_q$ and $(X_1, X_2) = (U_1, U_2)K$. Since K is invertible, we have:

$$2H(\mu) = H(U_1, U_2) = H(X_1, X_2) = H(X_1) + H(X_2|X_1) \tag{36}$$

and:

$$H(X_1) - H(\mu) = H(\mu) - H(X_2|X_1) \tag{37}$$

which is the entropy spread gained by using the transformation K. To maximize the spread, one may maximize $H(X_1) = H(aU_1 + cU_2)$ over the choice of a and c or simply $H(U_1 + cU_2)$ over the choice of c. Hence, the maximization problem depends only on the variable c (a can be set to one, and b, d only need to ensure that K is invertible), which leads to a kernel of the form $K = \left[\begin{smallmatrix} 1 & 0 \\ c & 1 \end{smallmatrix}\right]$. Note that to maximize the spread, one may alternatively minimize $H(X_2|X_1) = H(U_2|U_1 + cU_2)$.

We consider in particular channels which are "additive noise", in which case, one can equivalently study the "source" version of this problem as follows:

$$\lambda^*(\mu) = \arg \max_{\lambda \in \mathbb{F}_q} H(U_1 + \lambda U_2), \tag{38}$$

where U_1, U_2 are i.i.d. under μ. As discussed above, this is related with the previous problem by choosing:

$$K^*(W) = \begin{bmatrix} 1 & 0 \\ \lambda^*(\mu) & 1 \end{bmatrix},$$

where μ is the distribution of the noise of the channel W.

Our first observation about the optimal coefficients $\lambda^*(\mu)$ is in the context of $\mathbb{F}_3$ and follows immediately from Theorem 1.

Corollary 1. *For a probability distribution μ over $\mathbb{F}_3$,*

$$\lambda^*(\mu) = 2$$

if $\mu(1) \neq \mu(2)$, and $\lambda^*(\mu) = \{1, 2\}$ if $\mu(1) = \mu(2)$.

Figure 2 illustrates the improvements of the error probability of a polar code using the kernel $\left[\begin{smallmatrix}1&0\\2&1\end{smallmatrix}\right]$ instead of $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$ for a block length $n = 1024$ when the channel is an additive noise channel over $\mathbb{F}_3$ with noise distribution $\{0.7, 0.3, 0\}$.

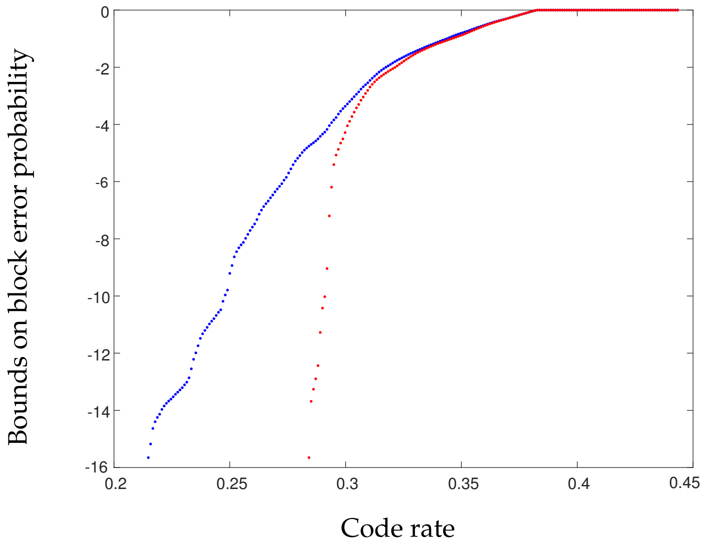

Figure 2. For an additive noise channel over $\mathbb{F}_3$ with noise distribution $\{0.7, 0.3, 0\}$, the block error probability (in $\log_{10}$ scale) of a polar code with block length of $n = 1024$ is plotted against the rate of the code. The red curve (lower curve) is for the polar code using the two-optimal kernel, whereas the blue curve is for the polar code using the original kernel.

When μ is over $\mathbb{F}_q$ with $q \geq 5$, $\lambda^*(\mu)$ varies with μ. For example, one can check numerically that for the distribution $\{0.8, 0.1, 0.1, 0, 0\}$, we have $\lambda^* = 4$, whereas for the distribution $\{0.7, 0.2, 0.1, 0, 0\}$, we have $\lambda^* = \{2, 3\}$. Thus, finding a solution to the problem of determining $\lambda^*(\mu)$ for general probability distributions μ on $\mathbb{F}_q$ seems not so easy. Nonetheless, for a certain class of probability distributions μ, we can identify $\lambda^*(\mu)$ explicitly using the following observation.

Proposition 1. *Let μ be a probability distribution over $\mathbb{F}_q$ with support S_μ. If there exists $\gamma \in \mathbb{F}_q$ such that:*

$$|S_\mu + \gamma S_\mu| = |S_\mu|^2 \tag{39}$$

then:

$$H(U_2|U_1 + \gamma U_2) = 0 \tag{40}$$

where U_1, U_2 are i.i.d. under μ.

Proof. The condition $|S_\mu + \gamma S_\mu| = |S_\mu|^2$ ensures that knowing $u_1 + \gamma u_2$ with $u_1, u_2 \in S_\mu$ allows one to exactly recover both u_1 and u_2. □

Remark 9. *The condition on the support could be simplified, but as such, it makes the conclusion of Proposition 1 immediate. Also note that γ such that $H(U_2|U_1 + \gamma U_2) = 0$ is clearly optimal to maximize the spread, i.e., it maximizes $H(U_1 + \gamma U_2)$.*

Let us consider some examples of distributions satisfying (39):

1. Let μ over $\mathbb{F}_5$ be such that $S_\mu = \{0, 1\}$. Picking $\gamma = 2$, one obtains $2S_\mu = \{0, 2\}$ and $S_\mu + 2S_\mu = \{0, 1, 2, 3\}$, and (39) is verified. In this case, using $\gamma = 1$ can only provide a strictly smaller spread since it will not set $H(U_2|U_1 + \gamma U_2) = 0$. It is hence better to use the two-optimal kernel

$\left[\begin{smallmatrix}1&0\\2&1\end{smallmatrix}\right]$ rather than the original kernel $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$. As illustrated in Figure 3, this leads to significant improvements in the error probability at finite block length. Also note that a channel with noise μ satisfying (39) has positive zero-error capacity, which is captured by the two-optimal kernel as shown with the rapid drop of the error probability (it is zero at low enough rates since half of the synthesized channels have noise entropy exactly zero). If μ is close to a distribution satisfying (39), the error probability can also be significantly improved with respect to the original kernel $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$.

2. Over $\mathbb{F}_{11}$, let μ be such that $S_\mu = \{0, 1, 2\}$. Picking $\gamma = 2$, one obtains $2S_\mu = \{0, 2, 4\}$, and (39) does not hold. However, picking $\gamma = 3$ leads to $3S_\mu = \{0, 3, 6\}$, and (39) holds. Therefore, the choice of γ varies with respect to q.
3. Over general $\mathbb{F}_q$, let $k = \lfloor\sqrt{q-1}\rfloor$. If $S_\mu = \{0, 1, \ldots, k-1\}$, we can see that $\gamma = k$ will satisfy (39).

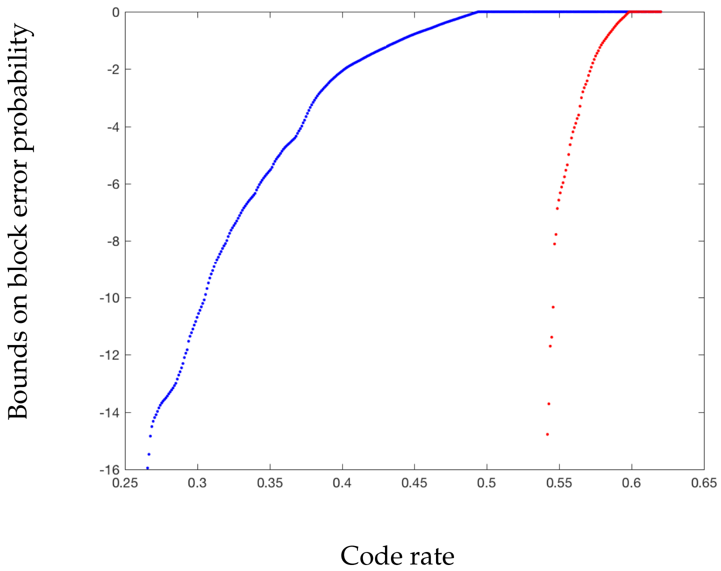

Figure 3. For an additive noise channel over $\mathbb{F}_5$ with noise distribution $\{0.7, 0.3, 0, 0, 0\}$, the block error probability (in $\log_{10}$ scale) of a polar code with block length of $n = 1024$ is plotted against the rate of the code. The red curve (lower curve) is for the polar code using the two-optimal kernel, whereas the blue curve is for the polar code using the original kernel.

In conclusion, we have shown that over $\mathbb{F}_q$, the martingale spread can be significantly enlarged by using two-optimal kernels rather than the original kernel $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$. Moreover, we have observed that this can lead to significant improvements on the error probability of polar codes, even at low block length ($n = 1024$). For additive noise channels, while the improvement is significant when the noise distribution is concentrated on a "small" support, the improvement may not be as significant for distributions that are more more spread out.

Acknowledgments: The authors would like to thank Emre Telatar for helpful comments and stimulating discussions. This work was supported in part by the U.S. National Science Foundation through the grants CCF-1319299, DMS-1409504 (CAREER) and CCF-1346564.

Author Contributions: All the authors contributed equally to the research and writing of the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Madiman, M.; Tetali, P. Information inequalities for joint distributions, with interpretations and applications. *IEEE Trans. Inf. Theory* **2010**, *56*, 2699–2713.
2. Tao, T.; Vu, V. Additive combinatorics. In *Cambridge Studies in Advanced Mathematics*; Cambridge University Press: Cambridge, UK, 2006; Volume 105, pp. xviii, 512.

3. Ruzsa, I.Z. Sumsets and Entropy. *Random Struct. Algoritms* **2009**, *34*, 1–10.
4. Madiman, M.; Marcus, A.; Tetali, P. Information-theoretic inequalities in additive combinatorics. In Proceedings of the 2010 IEEE Information Theory Workshop on Information Theory, Cairo, Egypt, 6–8 January 2010.
5. Madiman, M.; Marcus, A.; Tetali, P. Entropy and set cardinality inequalities for partition-determined functions. *Random Struct. Algoritms* **2012**, *40*, 399–424.
6. Tao, T. Sumset and inverse sumset theory for Shannon entropy. *Comb. Probab. Comput.* **2010**, *19*, 603–639.
7. Haghighatshoar, S.; Abbe, E.; Telatar, E. A new entropy power inequality for integer-valued random variables. *IEEE Trans. Inf. Theory* **2014**, *60*, 3787–3796.
8. Jog, V.; Anantharam, V. The entropy power inequality and Mrs. Gerber's lemma for groups of order 2^n. *IEEE Trans. Inf. Theory* **2014**, *60*, 3773–3786.
9. Wang, L.; Woo, J.O.; Madiman, M. A lower bound on the Rényi entropy of convolutions in the integers. In Proceedings of the 2014 IEEE International Symposium on Information Theory (ISIT), Honolulu, HI, USA, 29 June–4 July 2014; pp. 2829–2833.
10. Woo, J.O.; Madiman, M. A discrete entropy power inequality for uniform distributions. In Proceedings of the 2015 IEEE International Symposium on Information Theory (ISIT), Hongkong, China, 14–19 June 2015.
11. Madiman, M.; Kontoyiannis, I. The Entropies of the Sum and the Difference of Two IID Random Variables are Not Too Different. In Proceedings of the 2010 IEEE International Symposium on Information Theory Proceedings (ISIT), Austin, TX, USA, 13–18 June 2010.
12. Lapidoth, A.; Pete, G. On the Entropy of the Sum and of the Difference of Two Independent Random Variables. In Proceedings of the IEEE 25th Convention of Electrical and Electronics Engineers in Israel, Eilat, Israel, 3–5 December 2008.
13. Madiman, M. On the entropy of sums. In Proceedings of the 2008 IEEE Information Theory Workshop, Porto, Portugal, 5–9 May 2008; pp. 303–307.
14. Cohen, A.S.; Zamir, R. Entropy amplification property and the loss for writing on dirty paper. *IEEE Trans. Inf. Theory* **2008**, *54*, 1477–1487.
15. Etkin, R.H.; Ordentlich, E. The degrees-of-freedom of the *K*-user Gaussian interference channel is discontinuous at rational channel coefficients. *IEEE Trans. Inf. Theory* **2009**, *55*, 4932–4946.
16. Wu, Y.; Shamai, S.S.; Verdú, S. Information dimension and the degrees of freedom of the interference channel. *IEEE Trans. Inf. Theory* **2015**, *61*, 256–279.
17. Stotz, D.; Bölcskei, H. Degrees of freedom in vector interference channels. *IEEE Trans. Inf. Theory* **2016**, *62*, 4172–4197.
18. Kontoyiannis, I.; Madiman, M. Sumset and Inverse Sumset Inequalities for Differential Entropy and Mutual Information. *IEEE Trans. Inf. Theory* **2014**, *60*, 4503–4514.
19. Madiman, M.; Kontoyiannis, I. Entropy bounds on abelian groups and the Ruzsa divergence. *arXiv* **2015**, arXiv:1508.04089.
20. Bobkov, S.; Madiman, M. Dimensional behaviour of entropy and information. *C. R. Math.* **2011**, *349*, 201–204.
21. Madiman, M.; Melbourne, J.; Xu, P. Forward and reverse entropy power inequalities in convex geometry. *arXiv* **2016**, arXiv:1604.04225.
22. Madiman, M.; Ghassemi, F. Combinatorial entropy power inequalities: A preliminary study of the Stam region. *arXiv* **2017**, arXiv:1704.01177.
23. Fradelizi, M.; Madiman, M.; Marsiglietti, A.; Zvavitch, A. On the monotonicity of Minkowski sums towards convexity. *arXiv* **2017**, arXiv:1704.05486.
24. Harremoës, P.; Johnson, O.; Kontoyiannis, I. Thinning, entropy, and the law of thin numbers. *IEEE Trans. Inf. Theory* **2010**, *56*, 4228–4244.
25. Johnson, O.; Yu, Y. Monotonicity, thinning, and discrete versions of the entropy power inequality. *IEEE Trans. Inf. Theory* **2010**, *56*, 5387–5395.
26. Barbour, A.D.; Johnson, O.; Kontoyiannis, I.; Madiman, M. Compound Poisson approximation via information functionals. *Electron. J. Probab.* **2010**, *15*, 1344–1368.
27. Johnson, O.; Kontoyiannis, I.; Madiman, M. Log-concavity, ultra-log-concavity, and a maximum entropy property of discrete compound Poisson measures. *Discret. Appl. Math.* **2013**, *161*, 1232–1250.
28. Croft, H.T. *Research Problems*; Mimeographed Notes: Cambridge, UK, August 1967.
29. Macdonald, S.O.; Street, A.P. On Conway's conjecture for integer sets. *Bull. Aust. Math. Soc.* **1973**, *8*, 355–358.
30. Marica, J. On a conjecture of Conway. *Can. Math. Bull.* **1969**, *12*, 233–234.

31. Nathanson, M.B. Problems in additive number theory. I. In *Additive Combinatorics, CRM Proceedings Lecture Notes*; American Mathematical Society: Providence, RI, USA, 2007; Volume 43, pp. 263–270.
32. Stein, S.K. The cardinalities of $A + A$ and $A - A$. *Can. Math. Bull.* **1973**, *16*, 343–345.
33. Piccard, S. *Sur des Ensembles Parfaits*; Mém. University Neuchâtel, Secrétariat de l'Université: Neuchâtel, Switzerland, 1942; Volume 16, p. 172. (In French)
34. Martin, G.; O'Bryant, K. Many sets have more sums than differences. In *Additive Combinatorics, CRM Proceedings Lecture Notes*; American Mathematical Society: Providence, RI, USA, 2007; Volume 43, pp. 287–305.
35. Hegarty, P.; Miller, S.J. When almost all sets are difference dominated. *Random Struct. Algoritms* **2009**, *35*, 118–136.
36. Zhao, Y. Counting MSTD sets in finite abelian groups. *J. Number Theory* **2010**, *130*, 2308–2322.
37. Mossel, E. Gaussian bounds for noise correlation of functions. *Geom. Funct. Anal.* **2010**, *19*, 1713–1756.
38. Zhao, Y. Constructing MSTD sets using bidirectional ballot sequences. *J. Number Theory* **2010**, *130*, 1212–1220.
39. Stein, E.M.; Shakarchi, R. *Fourier Analysis: An Introduction Princeton Lectures in Analysis*; Princeton University Press: Princeton, NJ, USA, 2003; Volume 1, pp. xvi, 311.
40. Hegarty, P.V. Some explicit constructions of sets with more sums than differences. *Acta Arith.* **2007**, *130*, 61–77.
41. Ruzsa, I.Z. On the number of sums and differences. *Acta Math. Hung.* **1992**, *59*, 439–447.
42. Nathanson, M.B. Sets with more sums than differences. *Integers* **2007**, *7*, 1–24.
43. Li, J.; Madiman, M. A combinatorial approach to small ball inequalities for sums and differences. *arXiv* **2016**, arXiv:1601.03927.
44. Ruzsa, I.Z. On the cardinality of $A + A$ and $A - A$. In *Combinatorics (Proc. Fifth Hungarian Colloq., Keszthely, 1976), Vol. II*; János Bolyai Mathematical Society: Amsterdam, The Netherlands, 1978; Volume 18, pp. 933–938.
45. Pigarev, V.P.; Freĭman, G.A. The Relation between the Invariants R and T. In *Number-Theoretic Studies in the Markov Spectrum and in the Structural Theory of Set Addition (Russian)*; Kalinin Gos. University: Moscow, Russian, 1973; pp. 172–174.
46. Arıkan, E. Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Trans. Inf. Theory* **2009**, *55*, 3051–3073.
47. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2006; pp. xxiv, 748.
48. Roth, R.M. *Introduction to Coding Theory;* Cambridge University Press: Cambridge, UK, 2006.
49. Park, W.; Barg, A. Polar codes for q-ary channels, $q = 2^r$. *IEEE Trans. Inf. Theory* **2013**, *59*, 955–969.
50. Sahebi, A.G.; Pradhan, S.S. Multilevel channel polarization for arbitrary discrete memoryless channels. *IEEE Trans. Inf. Theory* **2013**, *59*, 7839–7857.
51. Mori, R.; Tanaka, T. Non-Binary Polar Codes using Reed-Solomon Codes and Algebraic Geometry Codes. In Proceedings of the 2010 IEEE Information Theory Workshop (ITW 2010), Dublin, Ireland, 30 August–3 September 2010.
52. Arıkan, E. Source polarization. In Proceedings of the 2010 IEEE International Symposium on Information Theory Proceedings (ISIT), Austin, TX, USA, 13–18 June 2010; pp. 899–903.
53. Abbe, E. Randomness and dependencies extraction via polarization. In Proceedings of the 2011 Information Theory and Applications Workshop (ITA), La Jolla, CA, USA, 6–11 February 2011; pp. 1–7.
54. Arıkan, E.; Telatar, E. On the rate of channel polarization. In Proceedings of the 2009 IEEE International Symposium on Information Theory (ISIT 2009), Seoul, Korea, 28 June–3 July 2009; pp. 1493–1495.
55. Abbe, E. Polar martingale of maximal spread. In Proceedings of the International Zurich Seminar, Zurich, Switzerland, 29 February–2 March 2012.
56. Hassani, S.H.; Alishahi, K.; Urbanke, R.L. Finite-length scaling for polar codes. *IEEE Trans. Inf. Theory* **2014**, *60*, 5875–5898.
57. Mondelli, M.; Hassani, S.H.; Urbanke, R.L. Unified Scaling of Polar Codes: Error Exponent, Scaling Exponent, Moderate Deviations, and Error Floors. *IEEE Trans. Inf. Theory* **2016**, *62*, 6698–6712.
58. Hassani, H. Polarization and Spatial Coupling: Two Techniques to Boost Performance. Ph.D. Thesis, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2013.
59. Korada, S.B.; Şaşoğlu, E.; Urbanke, R. Polar codes: Characterization of exponent, bounds, and constructions. *IEEE Trans. Inf. Theory* **2010**, *56*, 6253–6264.

60. Fazeli, A.; Vardy, A. On the Scaling Exponent of Binary Polarization Kernels. In Proceedings of the 2014 IEEE 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 30 September–3 October 2014; pp. 797–804.
61. Pfister, H.D.; Urbanke, R. Near-Optimal Finite-Length Scaling for Polar Codes over Large Alphabets. In Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain, 10–15 July 2016; pp. 215–219.